DOCUMENT RESUME

ED 386 493 TM 024 062

AUTHOR Zwick, Rebecca; And Others
TITLE Assessing Differential Item Functioning in Performance Tests.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-93-14
PUB DATE Mar 93
NOTE 45p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Educational Assessment; *Item Bias; *Multiple Choice Tests; Test Items; Test Reliability; Test Validity
IDENTIFIERS *Mantel Haenszel Procedure; *Performance Based Evaluation; Polytomous Variables

ABSTRACT

Although the belief has been expressed that performance assessments are intrinsically more fair than multiple-choice measures, some forms of performance assessment may in fact be more likely than conventional tests to tap construct-irrelevant factors. As performance assessment grows in popularity, it will be increasingly important to monitor the validity and fairness of alternative item types. The assessment of differential item functioning (DIF), as one component of this evaluation, can be helpful in investigating the effect on subpopulations of the introduction of performance tasks. Developing a DIF analysis strategy for performance measures requires decisions as to how the matching variable should be defined and how the analysis procedure should accommodate polytomous responses. In this study, two inferential procedures, extensions of the Mantel-Haenszel procedure, and two types of descriptive summaries that may be useful in assessing DIF in performance measures were explored and applied to simulated data. All the investigated statistical methods appear to be worthy of further study. Nine tables present analysis results. (Contains 32 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-93-14

ED 386 493

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ASSESSING DIFFERENTIAL ITEM FUNCTIONING IN PERFORMANCE TESTS

Rebecca Zwick
John R. Donoghue
Angela Grima

Educational Testing Service
Princeton, New Jersey
March 1993

TM024062

BEST COPY AVAILABLE

Assessing Differential Item Functioning in Performance Tests

Rebecca Zwick, John R. Donoghue, and Angela Grima

Educational Testing Service

We thank Paul Holland and Neil Dorans for their helpful reviews. A shorter version of this paper will appear in Journal of Educational Measurement.

Copyright © 1993. Educational Testing Service. All rights reserved.

Abstract

Although the belief has been expressed that performance assessments are intrinsically more fair than multiple-choice measures, some forms of performance assessment may in fact be more likely than conventional tests to tap construct-irrelevant factors. As performance assessment grows in popularity, it will be increasingly important to monitor the validity and fairness of alternative item types. The assessment of differential item functioning (DIF), as one component of this evaluation, can be helpful in investigating the effect on subpopulations of the introduction of performance tasks. Developing a DIF analysis strategy for performance measures requires decisions as to how the matching variable should be defined and how the analysis procedure should accommodate polytomous responses. In this study, two inferential procedures and two types of descriptive summaries that may be useful in assessing DIF in performance measures were explored and applied to simulated data. All the investigated statistics appear to be worthy of further study.

The recent reemergence of performance assessment methods has been described as marking the end of a six-decade era of educational testing and the beginning of a new era of assessment (Stiggins, 1991). Examples of nationwide testing and assessment programs that include polytomously scored performance tasks are the College Board Advanced Placement tests, the National Assessment of Educational Progress (NAEP) writing, reading, science, and mathematics assessments, the Praxis Series (successor to the NTE teacher assessment), and the ACT College Outcome Measures Program and Workkeys assessments. Among the challenges offered by the newfound popularity of performance assessment is the need for psychometric procedures for assessing validity, reliability, and item properties, including differential item functioning (DIF). This paper addresses the topic of DIF procedures for performance tasks.

Although the belief has been expressed that performance assessment provides a more equitable approach to testing than multiple-choice measures (see Hambleton & Murphy, 1992), some forms of performance assessment may in fact be more likely than conventional tests to tap construct-irrelevant factors. For example, some uses of portfolios may provide an advantage to students who have access to high-quality materials, good opportunities for study at home, and motivated and highly educated parents who can provide assistance. Thus, socioeconomic factors may play an undesirably large role in determining the quality of the portfolio. Also, when item responses are scored by raters who know the identity of each respondent or who can guess the respondent's gender or ethnicity, rater bias can occur. For example, if respondents tend to receive higher scores from raters of their own race (see Oppler, Campbell, Pulakos, & Borman, 1992), then respondents who are scored by same-race

raters will have an unfair advantage. There is evidence that adding performance sections to an existing test can, in some instances, lead to larger mean differences among ethnic groups (Dunbar, Koretz, & Hoover, 1991). Larger observed differences may occur either because groups are, in fact, more different in terms of the newly defined construct or because of construct-irrelevant factors. As performance assessment grows in popularity, it will be increasingly important to monitor the validity and fairness of alternative item types. DIF analysis procedures, as one component of this evaluation, can be helpful in investigating the effect on subpopulations of the introduction of performance tasks.

In this article, we first discuss DIF analysis for dichotomous items, focusing on the Mantel-Haenszel (MH; 1959) approach developed by Holland and Thayer (1988). We then discuss some general issues that must be addressed in developing DIF analysis methods for performance tasks. Following this, we describe two promising inferential procedures, along with summary statistics, for this type of DIF analysis. Next, we describe the simulation study we used to evaluate the DIF methods, as well as a feasibility study based on real data. Finally, we present our conclusions about the accuracy and utility of the methods.

Assessing DIF in Conventional Multiple-Choice Tests

Several methods are currently in use for assessing DIF in dichotomous items, including approaches based on item response theory (e.g., Thissen, Steinberg, & Wainer, 1988), logistic regression (Swaminathan & Rogers, 1990), standardization (Dorans & Kulick, 1986), and the MH procedure (Holland & Thayer, 1988). We focus here on the MH, which has gained wide acceptance as a useful method. In a typical application of the MH method,

observations are stratified on overall test score, S, and performance on an item of interest--the *studied item*--is compared for the *reference* and *focal* groups, conditional on score. The focal (F) group is the population of primary interest, such as women or members of ethnic minorities. The reference (R) group is the population to which F group item responses are to be compared. (The findings of Holland and Thayer, 1988; Zwick, 1990; and Donoghue, Holland, & Thayer, in press, support the inclusion of the studied item in the matching variable in the dichotomous case. Inclusion of the studied item in the matching variable in both the dichotomous and polytomous cases is discussed in more detail in subsequent sections.) Let X_i designate the score on the item of interest, where a score of 1 indicates a correct answer and a score of 0 indicates an incorrect answer. G is a group membership variable, where $G = F$ or R. Let s_k represent the score in the k^{th} level of the matching variable, $k = 1, 2, ..., K$. The MH approach approximates the uniformly most powerful unbiased test of the hypothesis

$$H_0: \frac{\frac{P(X_i = 1 | S = s_k, G = R)}{P(X_i = 0 | S = s_k, G = R)}}{\frac{P(X_i = 1 | S = s_k, G = F)}{P(X_i = 0 | S = s_k, G = F)}} = \alpha = 1, \; k = 1, 2, ..., K . \qquad [1]$$

versus the alternative,

$$H_1: \frac{\frac{P(X_i = 1 | S = s_k, G = R)}{P(X_i = 0 | S = s_k, G = R)}}{\frac{P(X_i = 1 | S = s_k, G = F)}{P(X_i = 0 | S = s_k, G = F)}} = \alpha, \; \alpha \neq 1, \; k = 1, 2, ..., K. \qquad [2]$$

(Holland & Thayer, 1988).

Assume that the data are organized as a $2 \times 2 \times K$ contingency table. Within the k^{th} level of the matching variable, A_k and C_k are the numbers of examinees in the R and F groups, respectively, who answer the item correctly, B_k and D_k are the numbers of examinees in the R and F groups, respectively, who answer the item incorrectly, n_{Rk} and n_{Fk} are the numbers of examinees in the R and F groups, m_{1k} and m_{0k} are the numbers of examinees who answer the item correctly and incorrectly, and n_{++k} is the total number of examinees.

The MH chi-square statistic, which has a chi-square distribution with one degree of freedom under H_0, is

$$MH\ \chi^2 = \frac{\left(\left|\sum_k A_k - \sum_k E(A_k)\right| - .5\right)^2}{\sum_k V(A_k)} \qquad [3]$$

where $E(A_k) = \frac{n_{Rk}\, m_{1k}}{n_{++k}}$ and $V(A_k) = \frac{n_{Rk}\, n_{Fk}\, m_{1k}\, m_{0k}}{n_{++k}^2 (n_{++k} - 1)}$. The subtraction of .5 in the numerator serves as a continuity correction.

The MH estimator of the common odds ratio, α, is

$$\hat{\alpha}_{MH} = \frac{\sum_k A_k\, D_k / n_{++k}}{\sum_k B_k\, C_k / n_{++k}} \qquad [4]$$

A measure of DIF that has become quite common is

$$MH\ D\text{-}DIF = -2.35\ ln\ \hat{\alpha}_{MH} \qquad [5]$$

This rescaling of the MH log odds ratio places it on the Educational Testing Service (ETS) delta scale of item difficulty (Holland & Thayer, 1985).

Assessing DIF in Performance Tasks

Whereas well-established DIF procedures exist for dichotomous items, there is not yet a consensus about how to conceptualize or measure DIF when the outcome of interest is not simply a 1 or 0, but a judgment of the quality of a writing portfolio, mathematical proof, or scientific experiment. Usually, such judgments are ordered, but in some instances, "scoring" an item may consist of determining which of several unordered solution strategies characterize the response. Developing a DIF analysis strategy for performance tasks requires that two major issues be addressed:

1. How should the matching variable be defined?
2. How should the analysis procedure accommodate the polytomous responses?

Each of these is addressed in turn. Other discussions of DIF methods for polytomous items are given by Dorans and Schmitt (1991), Hanson (1992), Miller, Spray, & Wilson (1992), Welch and Hoover (in press), and Zwick (1992a, 1992b).

Definition of the matching variable

Defining an appropriate matching variable for performance tasks is less than straightforward. A fundamental problem that arises is that an entire performance assessment may consist of very few tasks--possibly a single item. One option in this case is to match subjects using a measure that is external to the performance assessment, such as the score on

a multiple-choice test in the same subject area. This type of strategy is possible, for example, using data from the Advanced Placement examinations, which typically consist of multiple-choice and essay items in the same field of study. This matching strategy has been criticized, however, on the grounds that the multiple-choice portions of an exam may not assess the same attributes as the performance tasks (hence the need for the latter) and therefore may not be an appropriate basis for a matching variable. Even when responses to a substantial number of performance tasks are available for each examinee, the summation of these responses to form a matching variable may be objectionable because of concerns about dimensionality: A set of complex performance tasks may not be as highly interrelated as a typical set of multiple-choice items. A technique that may be helpful in constructing a matching variable in performance assessment is to make use of available demographic and attitude information, possibly in combination with scores on the set of performance tasks. A possible strategy for combining multiple measures into a single composite matching variable is propensity score matching (Rosenbaum and Rubin, 1985; see Zwick, 1992b for a DIF application).

Because we chose to focus our work on the DIF analysis methods themselves, we did not consider in detail the important issue of selecting a matching variable. In our simulation and real data applications, we matched on the score on a test that consisted of both dichotomous and polytomous items. We investigated several ways of computing this test score. First, we considered the effect of rescaling the scores on the polytomous items, which has been suggested as a way of ensuring that such items do not make a disproportionate contribution to the matching variable. We also evaluated the effect of excluding the studied

item from the matching variable. Some relevant theoretical considerations are given in the following section.

A Theoretical Perspective on Computation of the Matching Variable. In the dichotomous case, Holland and Thayer (1988) showed that under certain Rasch model conditions, identity of item response functions (IRFs) across groups for the studied item satisfies the MH null hypothesis (Equation 1) and a difference in IRFs across groups corresponds to the MH alternative hypothesis (Equation 2). The assumptions under which this finding holds are that (1) within each of the groups (R and F), the item response functions follow the Rasch model, (2) the matching variable is the number-right score based on all items, *including the studied item*, and (3) the items have the same IRFs for the two groups, with the possible exception of the studied item. Under these conditions, the odds ratios α_k in [1] and [2] are equal to $exp\ (b_F - b_R)$, where b_F and b_R are the Rasch item difficulties for the R and F groups, respectively.[1] The quantity $exp\ (b_F - b_R)$ is constant across all levels of the matching variable and is equal to one when the R and F groups have the same IRF. Zwick (1990) showed that the correspondence between IRF and MH definitions of DIF could not be assured to hold for a more general class of item response models, a finding that was confirmed empirically by the work of Donoghue, Holland, and Thayer (in press).

[1]This formulation applies to the case in which all item discrimination parameters are equal to unity and the scaling factor of 1.7 is not included in the item response function.

However, the theoretical findings of Zwick (1990) and the empirical results of Donoghue, Holland, and Thayer (in press) showed that inclusion of the studied item in the matching variable clearly improves the behavior of the MH odds ratio even when the data are generated by non-Rasch models (the three-parameter logistic [3PL] in Donoghue, Holland, & Thayer, and an even more general class of models in Zwick). Both studies showed that when no DIF is, in fact, present, departure of the MH odds ratio from its null value will be minimized by inclusion of the studied item. For DIF analysis of dichotomous items, these findings provide a strong justification for including the studied item, regardless of the presumed data generation model.

In attempting to extend the findings from the dichotomous case to polytomous items, it is useful to consider Masters' (1982) Rasch model for partial credit scoring. Assume that a particular task or item is scored on a scale with $T = M + 1$ ordered categories ranging from 0 to M. According to the partial credit model, the probability of receiving a score of x on item i for an examinee with proficiency θ is given by

$$\Pi_{xi}(\theta) \equiv \Pi_{xi} = \frac{\exp \sum_{m=0}^{x} (\theta - \delta_{mi})}{\sum_{p=0}^{M} \exp \sum_{m=0}^{p} (\theta - \delta_{mi})}, \quad x = 0, 1, ..., M. \qquad [6]$$

where δ_{mi} represents the difficulty of making the transition from category m - 1 to category m and $\sum_{m=0}^{0} (\theta - \delta_{mi})$ is assumed to be equal to zero by convention. Although the response categories are ordered, the difficulty parameters δ_{mi} need not be. For example, the transition from category 1 to category 2 may be more difficult than the transition from category 2 to

category 3. In the discourse below, the subscript R or F is appended to the δ_{mi} and the Π_{xi} from [6] to distinguish the values for the reference and focal groups.

The item response functions [6] for the partial credit model are of the same general form as the Rasch model for dichotomous items and Masters showed that the simple sum, S, of the item scores is a sufficient statistic for ability in the partial credit model, just as in the dichotomous Rasch model. (See Zwick, 1990 for a discussion of the relevance of sufficient statistics in this context.) The findings of Holland and Thayer (1988) can be generalized to the partial credit model as follows. Assume that all items follow the partial credit model and that all items, except possibly the studied item, are free of DIF, Consider two scores on the studied item, x_0 and $x_0 + q$. The odds ratio for these item scores, conditional on S, can be shown to be equal to

$$exp\left(\sum_{m=x_0+1}^{x_0+q} \delta_{miF} - \sum_{m=x_0+1}^{x_0+q} \delta_{miR}\right) \qquad [7]$$

As in the dichotomous case, this quantity is constant across levels of the matching variable (satisfying the analogue to [2]) and is equal to one (satisfying the analogue to [1]) if the ratio of the IRFs for the two scores is the same for the R and F groups; that is, if $\Pi_{(x_0+q)iR}/\Pi_{x_0iR} = \Pi_{(x_0+q)iF}/\Pi_{x_0iF}$. As in the dichotomous case, conditioning on the simple sum of the item scores leads to a desirable concordance, in the population, between a definition of DIF based on odds ratios and a definition based on IRFs.[2] Thus, there is some theoretical

[2] Note, however, that this definition of DIF applies to only the pair of categories in question. For example, suppose we have a three-category item with $\delta_{1iF} = \delta_{1iR} + c$, $\delta_{2iF} = \delta_{2iR}$, and $\delta_{3iF} = \delta_{3iR} - c$. Then for categories 2 and 0, the expression in [7] will be equal to unity

support for calculating the matching variable in the polytomous case by simply summing the scores on all the items, including the studied item.[3]

Extensions of MH Analysis Methods to Accommodate Polytomous Items

Two elaborations of the MH procedure appear to be promising for the assessment of DIF for polytomous items: Mantel (1963) proposed an extension of the MH procedure for ordered response categories which involves comparing the means for two groups, conditional on a matching variable. In addition, Mantel and Haenszel (1959; see also Somes, 1986) presented a generalized MH statistic (GMH) that is a direct extension of the ordinary MH to the case of $T > 2$ response categories. (The case of more than two groups was also considered by Mantel and Haenszel, but will not be addressed here.) The GMH statistic does not explicitly take into account the possible ordering of response categories; rather, it provides for a comparison of the two groups with respect to their entire response distributions, conditional on a matching variable. A general form of the MH statistic that subsumes both the Mantel and GMH procedures was given by Landis, Heyman, and Koch (1978; see also Agresti, 1990, p. 286). The goal of the current research was to investigate the utility of these extended MH methods and to determine whether interpretable indexes of DIF could be developed to supplement the results of the statistical hypothesis tests. The two MH

(i.e., DIF "balances out," as described in the simulation description below). However, for the category pairs 0 and 1 or 1 and 2, [7] will not equal unity.

[3]It is not clear how to include the studied item in the matching variable if the item has unordered scoring categories.

extensions were programmed by Donoghue for this study. Both are also included in SAS PROC FREQ (SAS Institute, Inc., 1990) under the heading of "Cochran-Mantel-Haenszel (CMH) Statistics."[4]

Mantel Approach for Ordered Response Categories. Mantel (1963) proposed a one-degree of freedom test of conditional association for the case of ordered response categories. Application of the method in the DIF context involves assigning index numbers to the response categories and then comparing the item means for members of the R and F groups who have been matched on a measure of proficiency. Welch and Hoover (in press) conducted a simulation study of the utility of this statistic for DIF analysis and Holland and Thayer (Holland, 1991) also investigated its use. The data are organized into a $2 \times T \times K$ contingency table, where T is the number of response categories and K is the number of levels of the matching variable. At each of the K levels, the data can be represented as a $2 \times T$ contingency table like that shown in Table 1.

Insert Table 1 about here

The values, $y_1, y_2, ..., y_T$ represent the T scores that can be obtained on the item (possibly, but not necessarily, the integers 1, 2, ..., T). The body of the table contains values of n_{Rk} and n_{Fk}, which denote the numbers of R and F group members, respectively, who are at the k^{th} level of the matching variable and received an item score of y_r. A "+" denotes summation over a

---

[4] In recognition of the fact that Cochran (1954) developed a procedure very similar to the Mantel-Haenszel test, some authors use the term, "Cochran-Mantel-Haenszel statistics," to refer to the MH test and its extensions.

particular index. For example, n_{F+k} denotes the total number of F group members at the k^{th} level.

The statistic proposed by Mantel, reformulated in the notation of this paper is

$$Mantel\ \chi^2 = \frac{\left(\sum_k F_k - \sum_k E\left(F_k\right)\right)^2}{\sum_k Var\left(F_k\right)}, \qquad [8]$$

where F_k, the sum of scores for the focal group at the k^{th} level of the matching variable, is defined as

$$F_k = \sum_i y_i\, n_{Fik}\,.$$

The expectation of F_k under the hypothesis of no association (H_0) is

$$E\left(F_k\right) = \frac{n_{F+k}}{n_{++k}} \sum_i y_i\, n_{+ik}\,,$$

and the variance of F_k under H_0 is

$$Var\left(F_k\right) = \frac{n_{R+k}\, n_{F+k}}{n_{++k}^2\left(n_{++k} - 1\right)} \left\{\left(n_{++k} \sum_i y_i^2\, n_{+ik}\right) - \left(\sum_i y_i\, n_{+ik}\right)^2\right\}. \qquad [9]$$

Under H_0, the statistic in [8] has a chi-squared distribution with one degree of freedom. In the case of dichotomous items with scores coded 0 and 1, this statistic is identical to the Mantel-Haenszel (1959) statistic without the continuity correction. In DIF

applications, rejection of H_0 suggests that members of the R and F groups who are similar in overall proficiency nevertheless tend to differ in their performance on the studied item.

Welch and Hoover (in press) used data generated from the Masters (1982) partial credit model to compare the Mantel procedure to two combined t-test procedures. All DIF was of the constant variety; that is, all transitions from a given item score category to the next highest category were assumed to be more difficult for the focal group, and the degree to which they were more difficult was constant across score categories. (See the simulation section below for further discussion of this type of DIF.) Ability distributions were assumed to be normal. The matching variable was the number right based on simulated responses to 71 dichotomous items. The factors that were varied across simulation conditions were the focal group mean, the R and F group sample sizes, and the magnitude of DIF in the studied item. Overall, the Mantel procedure was somewhat less likely to detect DIF than the other two methods. However, the Type I error rates for the combined t methods were more likely to exceed the nominal level than that of the Mantel method, which displayed excellent Type I error control.

GMH Statistic for Nominal Data. The test statistic for the GMH procedure is a multivariate generalization of [3] (see Somes, 1986). That is, A_k and $E(A_k)$ are now vectors of length T - 1, corresponding to (any) T - 1 of the T response categories, and $V(A_k)$ is a T - 1 by T - 1 covariance matrix.[5] In the notation of Table 1,

[5]If the marginal frequencies and any T - 1 of the cell frequencies are known (for a particular level of the matching variable), then the remaining frequencies are determined. That is, there are only T - 1 nonredundant values.

$$\underset{\sim}{A}'_k = \left(n_{R1k}, n_{R2k}, \cdots, n_{R(T-1)k}\right),$$

$$E\left(\underset{\sim}{A}'_k\right) = n_{R+k}\,\underset{\sim}{n}'_k \,/\, n_{++k},$$

$$\underset{\sim}{n}'_k = \left(n_{+1k}, n_{+2k}, \cdots, n_{+(T-1)k}\right),$$

$$\underset{\sim}{V}\left(\underset{\sim}{A}_k\right) = n_{R+k}\, n_{F+k}\left(\frac{n_{++k}\; diag\; \underset{\sim}{n}_k - \underset{\sim}{n}_k\,\underset{\sim}{n}'_k}{n^2_{++k}\left(n_{++k} - 1\right)}\right),$$

and $diag\,\underset{\sim}{n}_k$ is a $(T - 1) \times (T - 1)$ diagonal matrix with elements $\underset{\sim}{n}_k$. The test statistic is

$$GMH\ \chi^2 = \left[\Sigma\ \underset{\sim}{A}_k - \Sigma\ E\left(\underset{\sim}{A}_k\right)\right]' \left[\Sigma\ \underset{\sim}{V}\left(\underset{\sim}{A}_k\right)\right]^{-1} \left[\Sigma\ \underset{\sim}{A}_k - \Sigma\ E\left(\underset{\sim}{A}_k\right)\right]. \qquad [10]$$

The statistic in [10] has a chi-squared distribution with T - 1 degrees of freedom under H_0. For dichotomous variables, it reduces to the statistic in [3] without the continuity correction. As noted earlier, the GMH statistic does not explicitly take into account the possible ordering of response categories; instead, it provides for the comparison of the two groups in terms of their entire response distributions, rather than their means alone. The odds that focal group members will be assigned a particular score category can be compared to the odds for the reference group, conditional on the matching variable. This approach could be particularly useful in conducting distractor analysis, in assessing the propensity to omit, or in analyzing the occurrences of particular types of solution strategies, which may be unordered.

Descriptive Statistics for the Mantel-Haenszel Extensions. In addition to investigating hypothesis testing procedures, we examined the utility of various descriptive statistics analogous to [4] or [5] for the MH extensions. A possible summary statistic for the Mantel

(1963) approach is the standardized mean difference between the R and F groups proposed by Dorans and Schmitt (1991). This approach is analogous to the standardization statistic developed by Dorans and Kulick (1986) for the case of dichotomous items. The statistic compares the means of the R and F groups, adjusting for differences in the distribution of R and F group members across the values of the matching variable. The proposed statistic is labeled *SMD* for "standardized mean difference" in the present paper. Reformulated in the notation used here, it is defined as follows:

$$SMD = \sum_k p_{Fk}\, m_{Fk} - \sum_k p_{Fk}\, m_{Rk}\,, \qquad [11]$$

where $p_{Fk} = \frac{n_{F+k}}{n_{F++}}$ is the proportion of focal group members who are at the k^{th} level of the matching variable, $m_{Fk} = \frac{1}{n_{F+k}} \left(\sum_i y_i\, n_{Fik} \right)$ is the mean item score for the focal group at the k^{th} level, and $m_{Rk} = \frac{1}{n_{R+k}} \left(\sum_i y_i\, n_{Rik} \right)$ is the analogous value for the reference group. A standard error formula for *SMD* is derived in Zwick (1992a).

As in the Dorans and Kulick (1986) standardization statistic, the first term of [11] is just the grand mean of the item scores for the focal group. The second term is the mean item score for the reference group, "standardized" to give the reference group the same distribution across levels of the matching variable as that of the focal group.[6] A negative *SMD* value implies that, conditional on the matching variable, the F group has a lower mean item score

[6]In the *SMD* statistic, as in the *STD P-DIF* statistic, alternative weights can be substituted for the p_{Fk} values on the right-hand side of Equation 11.

than the R group. The calculation of *SMD* may help to make the results of Mantel's approach more interpretable to the user.

Developing a useful index of DIF to supplement the GMH χ^2 statistic appears to be a more difficult task. With two groups and T response categories, the GMH yields T - 1 independent estimates of conditional odds ratios analogous to the single conditional odds ratio associated with the ordinary MH. Suppose, for example, that an item has four possible score levels. The item could be an essay that is graded on a 0 to 3 scale or a mathematics problem with four unordered solution strategies, arbitrarily numbered 0 to 3. Consider R and F group members who are matched on overall test score. Regard an item score of 0 as a baseline and, for each of scores 1 through 3, consider the odds of receiving that score versus a score of 0. These odds can be estimated for both R and F groups, resulting in three conditional odds ratio estimates. The T - 1 conditional odds ratios can be estimated in various ways, including the similar statistics proposed by Mickey and Elashoff (1985) and Yanagawa and Fujii (1990) for biostatistical applications. The utility of these summaries in the DIF context was examined. Only the Yanagawa and Fujii statistics are described and reported here because they appeared to approximate more closely their theoretical values. (As described in the results section, both statistics produced inflated values.)

Let n_{Gtk} be the number of observations in group G, $G = R$ or F; score category t, $t = m$ or p; and level k of the matching variable, $k = 1, 2, \ldots, K$. Let n_{++k} be the total number of observations in level k, and $T = M + 1$ be the number of score categories, labeled 0 through M. The usual estimator of the conditional odds ratio for categories m and p is

$\hat{\alpha}_{mp} = \sum_{k=1}^{K} \frac{n_{Rmk} n_{Fpk}}{n_{++k}} \Big/ \sum_{k=1}^{K} \frac{n_{Rpk} n_{Fmk}}{n_{++k}}$. The Yanagawa and Fujii statistic for category p can be expressed as

$$YF_p = exp\left\{\left[\sum_{m=0}^{M} ln\left(\frac{\hat{\alpha}^{*}_{mp}}{\hat{\alpha}^{*}_{m0}}\right)\right]/T\right\}, \quad n \neq 0. \qquad [12]$$

where $\hat{\alpha}^{*}_{mp}$ is a more efficient estimator of α_{mp} obtained by applying an adjustment to $\hat{\alpha}_{mp}$ (Yanagawa & Fujii, 1990).[7] When there are only two score categories, 0 and 1, $YF_1 = \hat{\alpha}_{MH}$ (Equation 4).

In calculating these statistics, computational problems arose because of empty cells in the 2 x 2 x T tables. Therefore, the quantity .5 was added to each cell entry, as recommended by Gart (1962). More sophisticated smoothing techniques, such as those outlined in Bishop, Fienberg, and Holland (1975), could be used instead. The variances of these estimators were not computed, but are approximated by the variability across replications of the statistics in our simulation.

Simulation Study

To evaluate the performance of the methods when the properties of the data were known, a simulation was conducted and the two DIF methods were applied. Although the GMH method does not require that response categories be ordered, we focused on comparing the two methods when ordering did, in fact, exist. In each simulation condition, the total

[7]The adjustment, given in Yanagawa & Fujii, 1990, p. 746, Step 4, contains an error. Using the parameterization presented in the article, the correction term must be subtracted, rather than added, to duplicate the results reported by Yanagawa and Fujii.

number of test items was 25. Twenty-four of these (the *core* items) were used only in computing the matching variable; the 25th was the studied item. The factors that were varied across the simulation conditions were focal group ability distribution (2 levels) and characteristics of the studied item (27 levels). The item properties that defined the 27 levels included the difficulty parameters (δ_{mi}), pattern of DIF, and magnitude of DIF.

In designing our simulation, we chose the test length, mean difference between groups, and group sample sizes to resemble conditions that occur in actual DIF analyses at ETS. We chose DIF magnitudes that appeared to be realistic, based on previous research on DIF in dichotomous items. Welch and Hoover, in press, examined a wider range of DIF magnitudes, including much more extreme conditions than ours. In selecting patterns of DIF for our study, we included the constant DIF pattern examined by Welch and Hoover, as well as several other patterns that have been conjectured to be plausible.

The 2 x 27 = 54 simulation conditions were crossed with four ways of computing the matching variable. The four methods varied in terms of whether scores on polytomous items were rescaled in computing the matching variable (2 levels) and whether the studied item was included in the matching variable (2 levels). One hundred replications were conducted for each of the 54 x 4 = 216 combinations of simulation conditions with ways of forming the matching variable. The simulation design is summarized in Table 2. The properties of the simulated data are described in detail in the following sections.

Insert Table 2 about here

The statistics that were recorded for each studied item in each replication were the Mantel and GMH χ^2 statistics (Equations 8 and 10), the *SMD* statistic, (Equation 11), and the Yanagawa and Fujii (1990) statistic (Equation 12).

Reference and focal groups. In each condition, samples of 500 observations were drawn from the R and F distributions. The reference group distribution was standard normal (i.e., N(0,1)) in all conditions; the focal group distribution was either N(0,1) or N(-1,1).

Test specifications and item response models for core items. The first 24 of the 25 items in the test were used only in computing the matching variable. Items 1-20 were dichotomous and items 21-24 were four-category items. These 24 items were free of DIF.

For the dichotomous items, data were generated using the 3PL model. The item parameters were intended to be representative of values obtained in calibrations of operational test data. The *a* parameters were based on findings of Stocking (personal communication, April, 1992). She has found that empirical *a* parameters are well represented by the log-normal distribution, $a \sim exp(z)$, with $z \sim N(-.065, 0.13)$. Five *a* values were chosen (0.638, 0.741, 0.861, 1.000, 1.162), corresponding to mean, $z = \mu_z \pm 0.5\sigma_z$, and $\mu_z \pm 1.0\sigma_z$. The *b* parameters were spaced uniformly from -2.25 to 2.25, separated by 0.25, with an additional value of 0.0. The *b* parameters were then sorted and formed into groups of five. One member of each group was randomly matched to each of the five *a* parameters (i.e., a randomized block design was used). The lower asymptote *c* was set to 0.15 for all items. The parameter values for the dichotomous items (1-20) are given in the upper panel of Table 3.

For the polytomous items, data were generated according to the Masters (1982) partial credit model (Equation 6). Parameters for items 21-24 were selected from the item parameter estimates from an analysis of a developmental screening test reported in Masters' article. They are listed in the lower panel of Table 3. Parameters for the studied items are discussed in the next section.

Insert Table 3 about here

Characteristics of the Studied Item. As noted above, the first 24 items in each simulation condition were free of DIF and the 25th item played the role of the studied item, which could potentially have DIF. The studied item always had four categories. DIF was modeled by starting with a set of reference group difficulty parameters and then adding a value to one or more of the item difficulties. We included three sets of reference group parameters (designated A, B, and C), four patterns of DIF and two (nonzero) magnitudes of DIF (.1 and .25), resulting in 24 types of DIF items. In addition, for each of the three sets of reference group parameters, a null condition was included in which the studied item had the same parameters for the R and F groups. Therefore, the total number of studied items was 24 + 3 = 27. The three sets of reference group parameter values for the studied item, selected, as above, from Masters' (1982) analysis of a developmental screening test, are given in Table 4.

Insert Table 4 about here

The four patterns of DIF we considered were as follows:

(1) DIF that was constant across response categories: In this condition, all of the transitions from a given item score category to the next highest category were assumed to be more difficult for the focal group, and the degree to which they were more difficult was constant across score categories; that is, $\delta_{miF} = \delta_{miR} + c$, $m = 1, 2, 3$.

(2) DIF that balanced across score categories: In this condition, the transition from the lowest to the second category was more difficult for the focal group, while the transition from the third category to the highest category was easier for the focal group. The remaining transition was the same for the two groups. That is, $\delta_{1iF} = \delta_{1iR} + c$, $\delta_{2iF} = \delta_{2iR}$, and $\delta_{3iF} = \delta_{3iR} - c$.

(3) DIF that affected only the lower score categories: In this condition, the transition from the lowest to the second category was more difficult for the focal group. The remaining transitions were identical for the two groups ($\delta_{1iF} = \delta_{1iR} + c$, $\delta_{2iF} = \delta_{2iR}$, and $\delta_{3iF} = \delta_{3iR}$).

(4) DIF that affected only the higher score categories: In this condition, the transition from the third to the highest category was more difficult for the focal group. The remaining transitions were identical for the two groups ($\delta_{1iF} = \delta_{1iR}$, $\delta_{2iF} = \delta_{2iR}$, and $\delta_{3iF} = \delta_{3iR} + c$).

Item parameters for the 27 studied items for the reference and focal groups are given in Table 5.

Insert Table 5 about here

Computation of the matching variable. Two aspects of the computation of the matching variable were varied across conditions: The first was the relative weights given to the dichotomous and polytomous items. In one condition, no rescaling of item scores was performed, so that the score range for the dichotomous items was 0-1, whereas the range for polytomous items was 0-3. In the other condition, the score on the polytomous items was rescaled by dividing by 3; in this condition, the score range was 0-1 for both types of items. The other aspect of computation of the matching variable that was varied was inclusion of the studied item: The studied item either was or was not included in the summation of the item scores. As noted earlier, if all items on the test follow the Rasch or partial credit models, then (1) the simple sum of the item scores, including the studied item, is sufficient for ability, (2) the rescaled sum is not sufficient unless all items in the summation have the same number of score categories, and (3) the (simple or rescaled) sum of scores excluding the studied item is not sufficient.[8]

Results and Discussion

The results of the study are given in Tables 6-8. Because the variation of results across the three sets of reference group parameters did not appear to be meaningful, all results have been averaged over the three sets.

---

[8] The model we used for data generation departs slightly from the model for which the simple sum of scores is sufficient in that some of the items in the matching variable were 3PL, rather than Rasch or partial credit items. Robustness findings from the dichotomous case (Donoghue, Holland, & Thayer, in press) suggest this should make little difference. This was borne out by our own findings in Tables 6 and 7.

Table 6 presents chi-square results for the Mantel and GMH procedures in the null case, averaged over the three sets of reference group parameters. The three 2 x 2 tables for each procedure show the pairwise effects of inclusion of the studied item, rescaling of polytomous item scores, and focal group mean (-1 or 0, where 0 is the reference group mean). Each mean chi-square in Table 6 is an average over 600 replications. In the null case, the expected value of a chi-square statistic is its degrees of freedom (df = 1 for the Mantel and df = 3 for the GMH procedure). Table 6 contains one of the main findings of our study: Our results showed that, as in the dichotomous case, the studied item should be included in the matching variable. Also, scores on polytomous items should not be rescaled when calculating the matching variable. Ignoring these guidelines leads to an increase in Type I errors.

The uppermost table for each procedure, which is averaged over the two levels of focal group mean, shows clearly that the average chi-square is close to its expected value when the studied item is included and when rescaling is not applied. The standard deviations of the chi-squares (see the top line of Table 7) were also close to their theoretical values of $\sqrt{(2df)}$. The average chi-squares were inflated, corresponding to an increased Type I error rate, when rescaling was used and even more so when the studied item was excluded. These results are consistent with the dichotomous case and with the theoretical predictions based on the partial credit model.

Insert Tables 6-8 about here

The middle and bottom panels of Table 6 show the effects of rescaling and of inclusion separately for the two focal group means. The results show that when the R and F

means are the same, the use of inappropriate procedures to compute the matching variable does not have a detrimental effect, but when the means differ, the impact of the method of computation is substantial. Again, this is consistent with the dichotomous case, in which failure to include the studied item leads to distorted conclusions about DIF when the R and F groups have different distributions, but not when they have identical distributions. The effects of rescaling and inclusion are strikingly similar for the Mantel and GMH procedures.

Because the results for the null case showed clearly that the appropriate method of forming the matching variable is to take the simple sum of all item scores, including the studied item, results for the non-null cases are shown for only this method. More research is needed on other aspects of matching not considered here, such as multidimensional matching variables, external matching variables, and matching variables based on non-cognitive data.

Table 7 presents the chi-square results for the non-null conditions, with the null results from the appropriate cells of Table 6 included for comparison. The mean and standard deviation of the chi-square values are given, as well as the rate of rejection of the null hypothesis for $\alpha = .05$. The variation of chi-square results over focal group means was inconsistent; therefore, results were averaged over this factor (as well as over the three sets of reference group parameters). Each value in Table 7 represents the result of 600 replications.

Both the Mantel and GMH methods appear promising; the preference for one over the other may depend on the type of DIF that is of most interest. Both chi-square procedures had slightly conservative rejection rates in the null case. For the constant DIF conditions, which involved the largest shifts in means between R and F groups, the Mantel procedure was more powerful than the GMH, as expected. In the balanced DIF condition, the GMH was superior-

-dramatically so for DIF of magnitude .25, where the percent of rejection was about 25 for GMH, but only 4 for Mantel method. This, too, is consistent with expectation, since the Mantel method is sensitive to mean differences, which are expected to be minimal in this type of DIF, whereas the GMH is sensitive to between-group differences in the frequencies of any of the item scores. For the "shift low" and "shift high" conditions, in which DIF affected only one category transition, the procedures produced similar rejection rates. For all DIF patterns except the constant pattern, detection rates were very low (8% or less) for DIF of a magnitude of .1. For constant DIF of .1, the rejection rates were about 18% and 11%, respectively, for the Mantel and GMH methods. For DIF of magnitude .25, the rejection rates ranged from 13% to about 76%, with one exception--the Mantel method in the balanced condition, which had a rejection rate of 4%. (For the Mantel procedure, rejection rates in both the null and non-null cases were somewhat smaller than those obtained in the Welch and Hoover (in press) simulation under similar conditions. One possible reason for the difference in results is that Welch and Hoover did not include the studied item in the matching variable.)

Table 8 presents the means and standard deviations over 600 replications of the *SMD* and Yanagawa and Fujii (1990) statistics for the same DIF conditions included in Table 7. In computing the *YF* statistics, the lowest item score (0) was used as the reference category, as indicated in Equation 12, where α^{*}_{m0} appears in the denominator of the odds ratio. The standard deviations across replications provide an empirical estimate of the standard errors of these statistics. Under the null hypothesis, the *SMD* means were close to zero, as expected. Like the Mantel chi-square, the *SMD* statistic was sensitive mainly to constant DIF; it may provide a useful supplement to the chi-square when this type of DIF is of interest.

The T - 1 *YF* statistics per item provide a more detailed picture of the pattern of DIF, but several aspects of these statistics require further study. In the null case, the *YF* statistics should, in theory, be equal to unity. The observed *YF* means, however, were considerably inflated. To explore further the behavior of the *YF* statistics, the theoretical values of the statistics under the partial credit model were calculated using Equation 7 in combination with Equation 12. Although the partial credit model does not hold here for all items in the matching variable, these theoretical values do provide a reasonable guideline for evaluating the size of the statistics. The theoretical values are given in Table 9. Although the observed values were sometimes very close the their theoretical counterparts (as in the constant, .25 condition), the statistics were, in general, too large. The Mickey and Elashoff (1985) estimates of the conditional odds ratios, not shown here, also tended to have inflated values. This is especially surprising because the addition of .5 to each cell in the table, which was performed because of sparseness, causes the odds ratio estimates to be closer to their null values.

Insert Table 9 about here

An issue of interpretability that arises in computing the *YF* statistics is how to determine which item score category should be the reference category. The lowest score (0) was used in our study, but other choices may lead to more useful results. Also, although all T - 1 conditional odds ratios may provide valuable information, it would be desirable to have a single overall DIF index as in the case of dichotomous items. To make the *SMD* and *YF* summaries maximally useful for test developers and users, some rules for categorizing the

severity of DIF, analogous to the MH-based rules that have been developed for the dichotomous case, will be needed.

In addition to using simulated data, we also applied the DIF methods to male-female analyses of data collected in the 1990 NAEP reading and writing trend assessments (see Johnson & Allen, 1992). The sample consisted of about 2,000 eleventh grade examinees who were administered a combination of multiple-choice reading items, constructed response reading items, and constructed response writing items. The constructed response items were scored on an ordinal scale by trained readers. The main purpose of the NAEP analysis was to assess feasibility. We wanted to make sure that no computational problems arose when real data, rather than model-generated data were analyzed. No such problems occurred, and, in general, the two test statistics behaved as anticipated. As would be expected with ordered responses, it was not unusual to find items for which the Mantel procedure led to a statistically significant result, while the GMH did not. It was interesting, however, to find NAEP items in which the opposite occurred. In particular, on one essay item (scored on a 0 - 4 scale), the GMH procedure led to a statistically significant result for a comparison of males and females, while the Mantel procedure did not. Conditional on the matching variable, response distributions for the two groups were different on this item, but mean responses were similar, as in the balanced DIF condition in the simulated data. If the entire response distribution, rather than merely the means, is of interest, the GMH may be the procedure of choice even when the response categories are ordered.

As a final note, it must be remembered that DIF analyses, while they can be helpful in investigating the effect on subpopulations of the introduction of alternative item types, are

only one component of the extensive research that is needed on the validity and fairness of performance assessment.

References

Agresti, A. (1990). Categorical data analysis. New York: Wiley.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). Discrete multivariate analysis. Cambridge, MA: The MIT Press.

Cochran, W. G. (1954). Some methods of strengthening the common χ^2 tests. Biometrics, 10, 417-451.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (in press). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland and H. Wainer (eds.) Differential Item Functioning. Hillsdale, NJ: Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. Journal of Educational Measurement, 23, 355-368.

Dorans, N. J., & Schmitt, A. P. (1991). Constructed response and differential item functioning: A pragmatic approach. ETS Research Report 91-47. Princeton, NJ: Educational Testing Service.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. Applied Measurement in Education, 4, 289-303.

Gart, J. J. (1962). Approximate confidence limits for relative risks. Journal of the Royal Statistical Society, B, 26, 454-463.

Hambleton, R. K., & Murphy, E. (1992). A psychometric perspective on authentic measurement. Applied Measurement in Education, 5, 1-16.

Hanson, B. (1992). Comments on Miller and Spray paper. Internal memorandum, American College Testing Program.

Holland, P. W. (January 14, 1991). Item and DIF analyses for items with ordered responses. Internal memorandum, Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1985). An alternative definition of the ETS delta scale of item difficulty. ETS Research Report No. RR 85-43. Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), Test Validity, pp. 129-145. Hillsdale, NJ: Erlbaum.

Johnson, E. G., & Allen, N. L. (1992). The NAEP 1990 technical report (No. 21-TR-20). Washington, DC: National Center for Education Statistics.

Landis, J. R., Heyman, E. R., & Koch, G. G. (1978). Average partial association in three-way contingency tables: A review and discussion of alternative tests. International Statistical Review, 46, 237-254.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. Journal of the American Statistical Association, 58, 690-700.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

Mickey, R. M., & & Elashoff, R. M. (1985). A generalization of the Mantel-Haenszel estimator of partial association for 2 x J x K tables. Biometrics, 41, 623-635.

Miller, T., Spray, J., & Wilson, A. (July, 1992). A comparison of three methods for identifying nonuniform DIF in polytomously scored test items. Paper presented at the annual meeting of the Psychometric Society, Columbus, Ohio.

Oppler, S. H., Campbell, J. P., Pulakos, E. D., & Borman, W. C. (1992). Three approaches to the investigation of subgroup bias in performance measurement: Review, results, and conclusions. Journal of Applied Psychology, 77, 201-217.

Rosenbaum, P. R. & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician, 39, 33-38.

SAS Institute, Inc. (1990). SAS procedures guide, version 6, third edition. Cary, NC: Author.

Somes, G. W. (1986). The generalized Mantel-Haenszel statistic. The American Statistician, 40, 106-108.

Stiggins, R. J. (1991). Facing the challenges of a new era of educational assessment. Applied Measurement in Education, 4, 263-273.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression. Journal of Educational Measurement, 27, 361-370.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer and H. I. Braun (Eds.), Test Validity, pp. 147-169. Hillsdale, NJ: Erlbaum.

Welch, C., & Hoover, H. D. (in press). Procedures for extending item bias techniques to polytomously scored items. Applied Measurement in Education.

Yanagawa, T., & Fujii, Y. (1990). Homogeneity test with a generalized Mantel-Haenszel estimator for L2 x K contingency tables. Journal of the American Statistical Association, 85, 744-748.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? Journal of Educational Statistics, 15, 185-197.

Zwick, R. (1992a). Application of Mantel's score test to the analysis of differential item functioning for ordinal items. Under revision for ETS Research Report series.

Zwick, R. (1992b). Differential item functioning analysis for new modes of assessment (in collaboration with Donoghue, J., Grima, A., Holland, P., Thayer, D., Thomas, N., & Wingersky. M.). Presented at the annual meeting of the National Council of Measurement in Education, San Francisco.

Table 1

Data for the k^{th} Level of the Stratification Variable

	Item Score					
Group	y_1	y_2	y_3		y_T	Total
Reference	n_{R1k}	n_{R2k}	n_{R3k}	...	n_{RTk}	n_{R+k}
Focal	n_{F1k}	n_{F2k}	n_{F3k}	...	n_{FTk}	n_{F+k}
Total	n_{+1k}	n_{+2k}	n_{+3k}	...	n_{+Tk}	n_{++k}

Table 2

Factors Varied in the Study*
(Number and description of levels given in parentheses)

Simulation Conditions (54: 2 focal group means x 27 studied items)

- Focal Group Mean (2: -1 and 0)
- Characteristics of Studied Item (27: 3 sets reference group parameters x 9 DIF conditions)
  - Reference Group Parameters (3 sets: A, B, and C)
  - DIF Conditions (9: 8 non-null + 1 null:)
    - Non-Null (8: 4 patterns x 2 magnitudes)
      - Pattern of DIF (4: constant, balanced, shift low, shift high)
      - Magnitude of DIF (2: .1 and .25)
    - Null (1)

Method of Computing the Matching Variable (4: 2 inclusion conditions x 2 rescaling conditions)

- Inclusion of Studied Item (2: included and not included)
- Rescaling of Polytomous Item Scores (2: rescaled and not rescaled)

* 100 replications were conducted for each of the 54 x 4 = 216 combinations of simulation conditions with methods of computing the matching variable.

Table 3

Parameters for Items in the Core Test (Items 1-24)

Dichotomous Items			
Item	Parameter		
	a_i	b_i	c_i
1	0.741	-2.25	0.15
2	0.861	-2.00	0.15
3	1.162	-1.75	0.15
4	0.638	-1.50	0.15
5	1.000	-1.25	0.15
6	1.000	-1.00	0.15
7	1.162	-0.75	0.15
8	0.638	-0.50	0.15
9	0.741	-0.25	0.15
10	0.861	0.00	0.15
11	1.000	0.00	0.15
12	0.741	0.25	0.15
13	1.162	0.50	0.15
14	0.638	0.75	0.15
15	0.861	1.00	0.15
16	0.638	1.25	0.15
17	0.741	1.50	0.15
18	1.162	1.75	0.15
19	1.000	2.00	0.15
20	0.861	2.25	0.15
Polytomous Items			
Item	Parameter		
	δ_{i1}	δ_{i2}	δ_{i3}
21	-0.91	-0.93	1.29
22	-1.34	1.72	3.40
23	-1.76	0.09	0.19
24	-2.20	-1.33	-0.48

Table 4

Reference Group Item Parameters for Studied Items

Polytomous Items			
Item	Parameter		
	δ_{i1}	δ_{i2}	δ_{i3}
A	-0.91	0.98	0.21
B	-2.25	-1.80	1.66
C	-0.54	-2.11	0.74

Table 5
Reference and Focal Group Item Parameters for Studied Items

Type of DIF	Item	Reference Group			Focal Group		
		δ_{iR}	δ_{pR}	δ_{iR}	δ_{iF}	δ_{pF}	δ_{iF}
Null	A	-0.91	0.98	0.21	-0.91	0.98	0.21
	B	-2.25	-1.80	1.66	-2.25	-1.80	1.66
	C	-0.54	-2.11	0.74	-0.54	-2.11	0.74
Constant .10	A	-0.91	0.98	0.21	-0.81	1.08	0.31
	B	-2.25	-1.80	1.66	-2.15	-1.70	1.76
	C	-0.54	-2.11	0.74	-0.44	-2.01	0.84
Constant .25	A	-0.91	0.98	0.21	-0.66	1.23	0.46
	B	-2.25	-1.80	1.66	-2.00	-1.55	1.91
	C	-0.54	-2.11	0.74	-0.29	-1.86	0.99
Balanced .10	A	-0.91	0.98	0.21	-0.81	0.98	0.11
	B	-2.25	-1.80	1.66	-2.15	-1.80	1.56
	C	-0.54	-2.11	0.74	-0.44	-2.11	0.64
Balanced .25	A	-0.91	0.98	0.21	-0.66	0.98	-0.04
	B	-2.25	-1.80	1.66	-2.00	-1.80	1.41
	C	-0.54	-2.11	0.74	-0.29	-2.11	0.49
Shift Low .10	A	-0.91	0.98	0.21	-0.81	0.98	0.21
	B	-2.25	-1.80	1.66	-2.15	-1.80	1.66
	C	-0.54	-2.11	0.74	-0.44	-2.11	0.74
Shift Low .25	A	-0.91	0.98	0.21	-0.66	0.98	0.21
	B	-2.25	-1.80	1.66	-2.00	-1.80	1.66
	C	-0.54	-2.11	0.74	-0.29	-2.11	0.74
Shift High .10	A	-0.91	0.98	0.21	-0.91	0.98	0.31
	B	-2.25	-1.80	1.66	-2.25	-1.80	1.76
	C	-0.54	-2.11	0.74	-0.54	-2.11	0.84
Shift High .25	A	-0.91	0.98	0.21	-0.91	0.98	0.46
	B	-2.25	-1.80	1.66	-2.25	-1.80	1.91
	C	-0.54	-2.11	0.74	-0.54	-2.11	0.99

Table 6

Effects of Inclusion of Studied Item (I), Rescaling (R) of the Matching Variable, and Focal Group Mean (F) on the Means of the Mantel and GMH Chi-Square Statistics in the Null Case (Means Over 600 Replications)[a]

		Mantel χ^2 (df = 1)			GMH χ^2 (df = 3)		
		R			R		
		no	yes	average	no	yes	average
I	no	2.31	2.78	2.55	4.39	4.81	4.60
	yes	0.99	1.68	1.34	3.07	3.73	3.40
	average	1.65	2.23	1.94	3.73	4.27	4.00
		R			R		
		no	yes	average	no	yes	average
F	0	1.05	1.07	1.06	3.00	2.99	3.00
	-1	2.25	3.39	2.82	4.45	5.55	5.00
	average	1.65	2.23	1.94	3.73	4.27	4.00
		I			I		
		no	yes	average	no	yes	average
F	0	1.06	1.06	1.06	2.99	3.00	3.00
	-1	4.03	1.61	2.82	6.20	3.79	5.00
	average	2.55	1.34	1.94	4.60	3.40	4.00

[a]All 6 tables are averaged over the 3 sets of reference group parameters. The top 2 tables are averaged over focal group populations (F). The middle 2 tables are averaged over inclusion conditions (I). The bottom 2 tables are averaged over rescaling conditions (R).

Table 7

Chi-Square Results for Mantel and GMH Procedures
Means, Standard Deviations, and Rejection Rates (α = .05) for 600 Replications[a]

Type of DIF	Mantel (*df*=1)			GMH (*df*=3)		
	Mean χ^2	*SD* χ^2	% reject	Mean χ^2	*SD* χ^2	% reject
NULL	0.99	1.41	4.33	3.07	2.44	4.83
Constant .1	2.21	2.71	17.83	4.15	3.22	11.17
Constant .25	8.23	5.73	75.67	10.24	6.14	59.83
Balanced .1	1.00	1.32	4.50	3.42	2.64	6.83
Balanced .25	0.94	1.29	4.00	5.71	4.00	24.67
Shift Low .1	1.13	1.56	6.00	3.42	2.79	7.83
Shift Low .25	1.72	2.29	13.17	4.32	3.32	13.00
Shift High .1	1.31	1.76	7.33	3.35	2.73	8.00
Shift High .25	1.78	2.41	14.00	4.61	3.61	16.50

[a]Results are averaged over the 3 sets of reference group parameters and the 2 focal group distributions.

Table 8

Results for SMD and Yanagawa and Fujii Statistics
Means and Standard Deviations for 600 Replications[a]

Type of DIF	*SMD*		Yanagawa and Fujii					
			YF_1		YF_2		YF_3	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
NULL	0.014	0.054	1.090	0.243	1.152	0.282	1.231	0.353
Constant .1	-0.034	0.056	1.159	0.268	1.329	0.317	1.539	0.461
Constant .25	-0.110	0.060	1.284	0.305	1.650	0.400	2.151	0.627
Balanced .1	0.012	0.053	1.183	0.264	1.250	0.304	1.225	0.354
Balanced .25	0.008	0.053	1.338	0.288	1.419	0.336	1.229	0.362
Shift Low .1	-0.004	0.055	1.176	0.252	1.249	0.299	1.334	0.383
Shift Low .25	-0.026	0.059	1.306	0.295	1.377	0.333	1.477	0.444
Shift High .1	-0.005	0.057	1.113	0.241	1.152	0.261	1.386	0.406
Shift High .25	-0.022	0.058	1.084	0.232	1.133	0.269	1.527	0.446

[a]Results are averaged over the 3 sets of reference group parameters and the 2 focal group distributions.

Table 9

Theoretical Values for the Yanagawa and Fujii Statistic Under the Partial Credit Model*
(c = amount of DIF)

	YF_1			YF_2			YF_3		
		c			c			c	
Type of DIF	Formula	.1	.25	Formula	.1	.25	Formula	.1	.25
Constant	*exp c*	1.05	1.28	*exp* (2*c*)	1.22	1.65	*exp* (3*c*)	1.35	2.12
Balanced	*exp c*	1.05	1.28	*exp c*	1.05	1.28	1	1	1
Shift Low	*exp c*	1.05	1.28	*exp c*	1.05	1.28	*exp c*	1.05	1.28
Shift High	1	1	1	1	1	1	*exp c*	1.05	1.28

*For the null case, the theoretical values for YF_n, n = 1, 2, 3, are unity.